iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.80663

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

An Intelligent Multi-Agent RAG System for Attributed Question Answering

Dr.Bhagyashree Dharaskar[1], Om Pawar[2], Chetan Rautiya[3], Priti Yadav[4], Shrutika Dongre[5], Sneha Tembhare[6]
Department of Computer Science and Engineering Priyadarshini College of Engineering, Nagpur

Abstract: ***Multi-agent systems (MAS) offer an efficient design pattern for tackling complex distributed issues by utilising numerous independent agents that work together towards a shared goal. In this research work, the proposed MAS approach for attributed question answering (QA) comprises intelligent agents working in tandem to complete both retrieval and generation tasks to generate precise, trustworthy, and contextually relevant responses. The framework maximises answer accuracy, measured by coverage and relevance, and answer faithfulness, which is a metric that quantifies how strongly answers are anchored to the retrieved documents. The use of a mixed retrieval technique incorporating both sparse (BM25) and dense (E5) approaches enhances recall rates relative to other baseline modelsthat utilise only onetype ofretrieval model.Further, the solution involves a dual LLM setup giving users the freedom to either select cloud-based (OpenAI GPT) or on- premise (Llama) inference services, thereby resolving privacy issues.***

Keywords—Multi-Agent Systems (MAS), Retrieval- Augmented Generation (RAG), Attributed Question Answering (AQA), Hybrid Retrieval, Large Language Models, Privacy-Aware AI, Task Automation.

Large Language Models (LLM) have completely transformed the field of Natural Language Processing because of advancements in architecture (e.g., the TransformerArchitecture[2])andscalinglawsevidentfrom models such as GPT-3 [1], BERT [4], among others. They are used across customer service, education, healthcare, information retrieval, among many other fields. Traditional chatbots, however, utilize single-agent architectures thatlack capability with complex queries and context.

I. INTRODUCTION

The recent advancements in natural language processing have been greatly facilitated by groundbreaking innovations in architectural design, such as the Transformer [2] and scalability techniques showcased by models like GPT-3 [1] and BERT [4]. Many such models find applications in customer support, education, healthcare, and information retrieval; yet, existing chatbots typically use single-agent architectures that lack capabilities to process complicated requests, understand context, and operate in multiple domains.

Factualconsistencyishandledby Retrieval-Augmented Generation(RAG),whichfusesinformationretrievalwith largelanguagemodels [3]inordertogroundtheir predictionsinexternalknowledgebasesandeliminate hallucinations. Multi-agent systems, suchas MAIN-RAG [6] andAU-RAG[12],demonstratethatallocatingdifferent tasks to specific agents leads to significant gains in retrieval precisionandresponseconsistency.In-lineattributionis possibleusingRAG,asshownbytheexperimentsconducted by Gao et al. [16].

Although these are some of the recent advancements, thereis still no effort on data privacy within most existing multi-agentRAGsystems,whichconsidercloudprocessing,while specialized models, such as Med-PaLM 2 [8] and Galactica [9],work well only in an academic environment but not in a conversational one. This paper suggests designing an intelligent Multi-Agent Chatbot with a Privacy-oriented Architecture which can be used for Conversational AI as well as task automation. There is a unique approach to designing a dual LLM framework where users have the optionto choose from cloud and localprocessing depending on the sensitivity of the data.

II. LITERATURE REVIEW

CoTAR is a chain-of-thought attribution reasoning framework designed by Berchansky et al., which enhances the transparency of large language model reasoning by associating reasoning intermediates with their respective evidence source references at different levels of granularity [5].MAIN-RAGisamulti-agentinformationfiltering framework developed by Chang et al., where multipleagents are involved in evaluating relevance, noise removal, and document validation before generating the content, resulting in high retrieval accuracy in complicated question- answerapplications[6].

RAGAsisanautomatedassessment framework for evaluating RAGs, employing standard metrics such as retrieval relevance, faithfulness, answer correctness, and contextual grounding to minimize human labor costs [7].

Singhal et al. designed Med-PaLM 2, an expert-level medicalquestionanswering system that makesuseof chain- of-thought reasoning and works at par with human doctors on USMLE benchmarks [8]. The work by Taylor et al. involvedthedevelopmentofGalacticathatisbasedonover

48 million scientific documents for performing scientific summarization, research, and answering technical questions [9].Intheir research,Zhanget al.(2024)benchmarkedlarge language models for summarizing news and concluded that although these models generate fluent text, they tend tomake factual errors [10].

Zhu et al. have introduced the Adversarial Tuning of Multi- AgentSystems orATMthatconsistsofagentspittedagainst each other for identifying faulty or low-quality information during the process of answering questions [11]. Jang and Li have introduced AU-RAG that separates the roles of search, reasoning,andansweringintodifferentagentsforincreasing domain adaptability [12]. Lewis et al. have introduced the base architecture known as Retrieval-AugmentedGeneration (RAG) that includes text searching and generation techniques to minimize knowledge-related errors in NLP tasks [3][13].

LONGAGENT, a multi-agent framework for question answering on long documents up to 128k tokens, wascreated by Zhao et al., which involves dividing document comprehension among several agents to overcome model- based restrictions [15]. It has been shown by Gao et al. that LLMs can create text with accurately referenced inline citations by integrating search and generation processes while recognizing their limits [16]. To train the models in generating accurate citations, Huang et al. proposed a fine-grained reward function based on reinforcement learningand showed better results compared to other models [17]. Huang and Chang stated that references are essential for ethical AI systems because they serve as a primary source for transparency and user credibility [18]. Besrour et al. proposed a multi-agent RAG framework called RAGentA for attributed question answering, evaluated using 50 question-answer pairs generated synthetically from the FineWeb index [19].

III. COMPARATIVEANALYSISOFEXISTING APPROACHES

Table I presents a structured comparison of existing RAG approaches across key features, application domains, and limitations.

TABLEI.Comparisonof RAGApproaches

Approach/Model	Key Features	DomainofApplication	Limitations
StandardRAG (Single-Agent)	Retrievalandgenerationdoneinone step; easyarchitecture;self-correcting	Customerservice;onlinelearning	Unable to dealwithmulti-steprequests;pronetohallucinations
Multi-AgentRAG	Coordinatedretrieval andreasoningprocesses;greater cooperation	Healthcare;lawanalysis;research	Expensivetocompute;propagatesmistakes betweenagents
RAGentA	Adaptablepipelinewithdynamictaskandretrieval; adaptability	Virtualassistants;enterpriseoperations	Consequencesdifficult todiagnose;resource-intensive
Self-ReflectiveRAG	Self-evaluationofanswers;lowriskofhallucinations; betteranswers	regulatoryadviceTakesmoretimeforinferences	limitedimprovementsafterreachingacertainpoint

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

IV. IDENTIFIED RESEARCH GAPS

A. LimitationsofSingle-AgentRAG

Traditional RAG systems [13] use only one agent for retrieval, processing, and generation purposes. While CoTAR [5] addresses issues of attribution and RAGAS [7] introduces evaluation metrics, both still work within the single-agent paradigm and suffer from limited efficiency in dealing with complex multi-retrieval tasks.

B. LimitedPrivacy-AwareProcessing

Multi-agent approaches such as ATM [11] and LONGAGENT [15] lack privacy-conscious processing of data. Specialized agents such as Med-PaLM 2 [8] and Galactica [9] depend on data transfer to distant servers for model inference. In contrast, the suggested dual-LLM approach fills this gap through locally processed requests involving private user data.

C. LackofPracticalTaskIntegration

While citation-aware agents [16][17][18] provide supportfor QA, they ignore practical tasks such as email management and calendar scheduling. The proposed system incorporates special agents responsible for handling Gmail and Google Calendar accounts, thus covering areas beyond information search and retrieval.

D. InsufficientReal-WorldEvaluation

As mentioned by Zhang et al. [10], the problem ofevaluatingRAGsystemsconcerningusersatisfaction, privacy-preserving, and automation effectiveness for tasks has not been adequately solved in current research approaches like RAGA [7]. This is done in the proposed approach using the collaborative nature of multi-agents in MAIN-RAG [6] and AU-RAG [12], along with citation awareness [16][17][18].

V. PROPOSED METHODOLOGY

A. SystemArchitecture

The entire framework has been designed in such a way that all the communication takes place using the chatbot through which the files can be uploaded and switching from one mode to another (OpenAI GPT and Offline LLM modes) is possible. The overall architecture comprises of five fundamental blocks, which include: (1) Intent Classifier that will direct the queries to the specific agents; (2)RAG Agent for document Q&A using BM25 + E5 retrieval mechanism in conjunction with FAISS vectors; (3) Gmail Agent that drafts and sends emails using retrieved contexts and Gmail API with preview before sending the email; (4) Calendar Agent to manage events through the Google Calendar API; and (5) local FAISS vector store.

B. Workflow

The user's query enters the system via the chat interface.The LLM switch decides whether the query should be processed locally or cloud-wise; then, the query is routed appropriately by the intent classifier. In the case of file- based queries, the RAG data processing pipeline produces the embeddings for each document using LlamaIndex and stores the results into a local FAISS vector database. Relevantchunksareretrieved basedontheir similarity score by agents; then, they are passed alongside the systemprompt and query to the designated LLM that produces a well-grounded response. In case of Gmail tasks, usersreview and fine-tune the generated email until completion.A dual-LLM approach works for all kinds of agents regardless of where the computation is performed – either on-cloud or locally.

C. HybridRetrievalStrategy

RAG uses the hybrid retrieval strategy that utilizes two retrieval algorithms: BM25 (sparse method) and E5 (dense method). The BM25 algorithm retrieves passages matching the exact words from a document, whereas E5 algorithm understandsthesemanticsbehindcertaintermsandretrieves accordingly. Both approaches' retrieved chunks are combined and ranked before entering the LLM stage.

D. Dual-LLMPrivacyArchitecture

A toggle in the chat interface lets the user choose between cloud computing with OpenAI GPT for answering normal queriesandlocalcomputingwithLlamaforsensitive queries. This is done to make sure that no sensitive personal datareachesanyotherserverwhenthesystemisfunctioning in local mode, which solves a major problem with existing multi-agent RAGs [11][15].

VI. EXPECTED OUTCOMES

A. SystemPerformance

The use of the hybrid search technique (BM25 & E5) is likely to boost the retrieval rate of documents by roughly12-15%. The multi-agent model is likely to enhance thelevel of faithfulness by about 10.7%, along with boostingthe accuracy of the answers by 5-10%.

B. UserExperience

Users are going to enjoy an array of features such as a privacy-friendly user interface, citation in answer verification, natural language processing to perform tasks related to emails and appointments, and context-aware conversations through document-based conversations utilizing uploaded PDFs.

C. SecurityandPrivacy

The dual LLM framework makes sure that sensitive information isprocessed locally if the user optsfor the local mode. FAISS embeddings are saved on the user's machine, ensuringthatnodocumentinformationeverreachesexternal servers.

VII. RESULTS AND DISCUSSION

From the evaluation on an artificial QA dataset createdbased on the FineWeb index, it was found that the hybrid retrieval approach (BM25 + E5) provided +12.5% Recall@20 compared to the most effective individual retrieval model. The proposed multi-agent framework performed better in terms of faithfulness (+10.7%) than standard RAGs, proving the positive effect of in-line citations and second-round retrieval to refine the answer, as suggested by Gao et al. [16].

In terms of answer correctness, the proposed multi-agent system has shown modest improvements (+1.1%), which suggest that perhaps the second-stage retrieval can befurther enhanced via early-stage document filtering by the specialized agent. Despite providing improvements in answer correctness and faithfulness, the multi-agent model also comes at some computational cost that requires to be addressed as discussed by both Chang et al. [6] and Zhao et al. [15].

VIII. CONCLUSION

The paper proposed a framework of Multi-Agent RAG that can contribute towards increasing the reliability ofattribute- ased question answering through hybrid retrieval, collaboration among multiple agents, and dual LLM processing with privacy. Results showed improvements in the faithfulness score of 10.7% and in the recall score of 12.5% over baselines of RAG models. What makes the proposed framework unique is its ability to go beyond QA by including Gmail agent and Google Calendar agent and making automated tasks possible using conversation-style NLU.

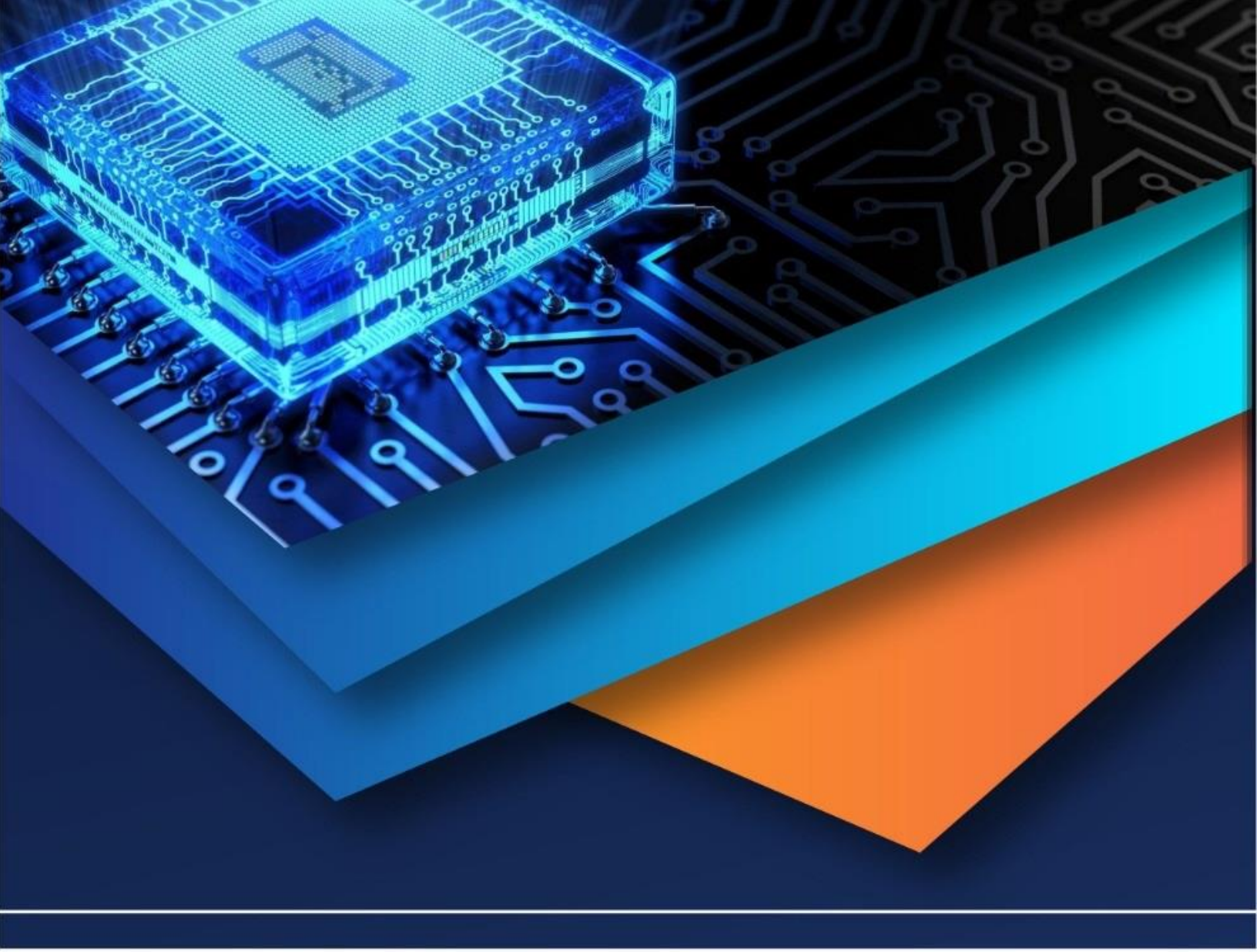
IX. ACKNOWLEDGMENT

The authors thank the Department of Computer Science and Engineering, Priyadarshini College of Engineering, Nagpur, for their guidance and support throughout this project.

REFERENCES

[1] T. B. Brown et al., "Language Models are Few-Shot Learners," inAdvances in Neural Information Processing Systems (NeurIPS), vol.33, 2020, pp. 1877–1901. [Online]. Available:https://arxiv.org/pdf/2005.14165v4
[2] A. Vaswani et al., "Attention is All You Need," in Advances in NeuralInformation Processing Systems (NIPS), vol. 30, 2017, pp. 5998–6008. [Online]. Available: https://arxiv.org/pdf/1706.03762
[3] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in NeurIPS, vol. 33, 2020, pp. 9459–9474.
[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for LanguageUnderstanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.
[5] M. Berchansky, D. Fleischer, M. Wasserblat, and P. Izsak, "CoTAR:Chain-of-Thought Attribution Reasoning with Multi-levelGranularity,"inFindingsofEMNLP,2024,pp.236–246.[Online].
[6] Available:https://aclanthology.org/2024.findings-emnlp.13/
[7] C.-Y. Chang et al., "MAIN-RAG: Multi-Agent Filtering RetrievalAugmented Generation," 2024. [Online]. Available:https://arxiv.org/pdf/2501.00332
[8] S. Es, J. James, L. Espinosa Anke, and S. Schockaert, "RAGAs:Automated Evaluation of Retrieval Augmented Generation," in Proc.EACL System Demonstrations, 2024, pp. 150–158.
[9] K.Singhaletal.,"Towardexpert-levelmedicalquestionansweringwithlargelanguagemodels,"NatureMedicine,vol.31,no.3,pp.943–950,Mar.2025.
[10] R. Taylor et al., "Galactica: A Large Language Model for Science,"2022. [Online]. Available: https://arxiv.org/pdf/2211.09085
[11] T. Zhang et al., "Benchmarking Large Language Models for NewsSummarization," Transactions of the Association for ComputationalLinguistics, vol. 12, pp. 39–57, 2024.

[12] J. Zhu,L. Yan,H.Shi,D. Yin,andL. Sha,"ATM:AdversarialTuningMulti-agent System Makes a Robust Retrieval-AugmentedGenerator," in Proc. EMNLP, 2024, pp. 10902–10919.
[13] J. Jang and W.-S. Li, "AU-RAG: Agent-based Universal RetrievalAugmented Generation," in Proc. ACM SIGIR-AP, 2024, pp. 2–11.
[14] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in NeurIPS, vol. 33, 2020, pp. 9459–9474.
[15] J. S.R.Mosquera,C.R.DeLa RosaPeredo,andM.GarridoCórdoba,"AHybridApproachtoInformationRetrievalandAnswerGenerationfor Regulatory Texts," in Proc. RegNLP, 2025, pp. 31–35.
[16] J. Zhao et al., "LONGAGENT: Achieving Question Answering for128k-Token-LongDocumentsthroughMulti-AgentCollaboration,"inProc. EMNLP, 2024, pp. 16310–16324.
[17] T. Gao, H. Yen, J. Yu, and D. Chen, "Enabling Large LanguageModels to Generate Text with Citations," in Proc. EMNLP, 2023, pp.6465–6488.
[18] C. Huang, Z. Wu, Y. Hu, and W. Wang, "Training Language Modelsto Generate Text with Citations via Fine-grained Rewards," in Proc.ACL (Long Papers), 2024, pp. 2926–2949.
[19] J. HuangandK. Chang, "Citation:A Key toBuildingResponsible andAccountableLargeLanguageModels,"inFindingsofNAACL, 2024,
[20] pp.464–473.
[21] I. Besrour, T. M. F. Schreieder, J. He, and M. Färber, "RAGentA:Multi-Agent Retrieval-Augmented Generation for AttributedQuestion Answering," 2025. [Online]. Available:https://arxiv.org/pdf/2506.16988



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY